Independence Testing for Time Series

by

Ronak Mehta

A thesis submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science and Engineering

Baltimore, Maryland
May, 2019

Abstract

Independence testing is a fundamental problem in statistical data analysis and machine learning settings. Before characterizing some predictive relationship between two modes of data (say, X and Y), a natural first question to ask is whether the phenomena are related at all. In statistical terms, we ask whether the random variables X and Y are independent. In the case that they are dependent, what geometry underlies their relationship? While there exist successful independence tests that operate on sets of independent and identically distributed (i.i.d) observations $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, corresponding procedures are limited for internally dependent data, such as time series. Procedures that exist often can only recognize linear relationships, unsuited for the complex nonlinear relationships present in real data. This work extends the independence testing problem to time series $\{(X_t, Y_t)\}$ processes, and addresses the unique challenges that come with estimation and testing among dependent data via a block permutation procedure. We address not only the existence of a relationship between two time series, but the spatial (geometric) and temporal nature of this relationship. Via simulations, we observe strong evidence of consistency in nonlinear settings. This initial work opens many doors for theoretical understanding as well as produces tools applicable to many real-world time series analysis problems.

Acknowledgments

This work would not be possible without: the dedicated guidance from Dr. Joshua T. Vogelstein and Dr. Cencheng Shen, the support from Hayden Helm, Ben Pedigo, Jaewon Chung, and Bijan Varjavand, the advice from Dr. Donniell Fishkind and Dr. Carey Priebe, and the Department of Applied Mathematics and Statistics at JHU.

Table of Contents

List of Figures

1 Introduction

In many data analysis and machine learning settings, a researcher might wish to determine the relationship between two jointly-observed phenomena, such as brain images and IQ scores, or the correspondence between stock prices in China and those in the United States. While further analysis can involve predicting the value of one phenomenon given an observation of the other or analyzing the geometry of the relationship, the first step is determining whether a discernible relationship exists. In statistical terms, we question whether the random variables representing the phenomena are independent (Vogelstein and Shen, 2019).

For many applications, the nature of this dependence can be highly non-linear, and linear dependence measures such as the Pearson's correlation coefficient may be insufficient (Vogelstein and Shen, 2019). An example is the phenomenon of volatility clustering in financial returns over time. These returns generally show no signs of linear dependence on past values, yet squared returns (a measure of financial volatility) tend to be highly correlated with previous values (Behrens and Sporns, 2012).

Multiscale graph correlation (`MGC`) is an existing independence testing procedure that is highly successful in terms of 1) applications to virtually any modality of data, 2) characterizing many types of geometric relationships, and 3) unmatched power at low sample sizes and high dimensionality. Other approaches include kernel-methods such as Hilbert-Schmidt Information Criterion (`HSIC`) (Gretton, 2005), and distance based methods such as distance

correlation (DCorr), of which MGC is an improvement. While Shen and Vogelstein, 2018 has shown distance and kernel-based methods to be equivalent, these procedures assume that observations from either modality are independent and identically-distributed. In the temporally-dependent time series setting, where data such as functional magnetic resonance images (fMRI), dynamically changing social networks, and the aforementioned financial index example are common, independence testing procedures are limited (Wang and Zhu, 2018). Researchers must resort to measures of linear dependence such as autocorrelation and crosscorrelation, ruling out potential nonlinear relationships (Wang and Zhu, 2018). The current challenge is to apply distance-based independence testing to time series (among other dependent data).

We propose cross-distance correlation (DCorr-X) and cross multiscale graph correlation (MGC-X), statistical independence tests for two multidimensional time series based on DCorr and MGC, respectively. We offer consistency results both linear and nonlinear settings, as well as an analysis neural connectivity via fMRI data. Additionally, our procedure estimates the time lag at which this dependence is maximized, further characterizing the temporal nature of the relationship. These methods expand the scope of traditional independence testing to complex, non-i.i.d. settings, accelerating research capabilities in neuroscience, econometrics, sociology, and many other fields.

2 Notation

Let $\mathbb{N}$ be the natural numbers $\{0, 1, 2, ...\}$, $\mathbb{Z}$ be the integers $\{..., -1, 0, 1, ...\}$, and $\mathbb{R}$ be the real line $(-\infty, \infty)$. Let F_X, F_Y, and $F_{(X,Y)}$ represent the marginal

and joint distributions of random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ on sample space Ω, respectively. Similarly, Let F_{X_t}, F_{Y_s}, and $F_{(X_t,Y_s)}$ represent the marginal and joint distributions of the time-indexed random variables $X_t : \Omega_t \to \mathcal{X}$ and $Y_s : \Omega_s \to \mathcal{Y}$ on sample spaces Ω_t and Ω_s at timesteps t and s. Assume $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ for $p, q \in \mathbb{N}$. Finally, let $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ represent the full, jointly-sampled time series, structured as a countably long list of observations (X_t, Y_t) indexed by t.

3 Problem Statement

Consider a strictly stationary time series $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$, with the observed sample $\{(X_1, Y_1), ..., (X_n, Y_n)\}$. Choose some $M \in \mathbb{N}$, the "maximum lag" hyperparameter. We wish to test the following hypothesis.

$$H_0 : F_{(X_t,Y_{t-j})} = F_{X_t} F_{Y_{t-j}} \text{ for each } j \in \{0, 1, ..., M\},\ t \in \mathbb{Z}$$

$$H_A : F_{(X_t,Y_{t-j})} \neq F_{X_t} F_{Y_{t-j}} \text{ for some } j \in \{0, 1, ..., M\}, t \in \mathbb{Z}$$

The null hypothesis implies that for any $(M + 1)$-length stretch in the time series, X_t is independent of past values Y_{t-j} spaced j timesteps away (including $j = 0$). A corresponding test also exists for whether Y_t is dependent on past values of X_t, by swapping the labels of each time series.

4 Preliminaries

4.1 Distance covariance and correlation

Consider random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, with finite first and second moments. The *distance covariance* function is defined as the positive square root of:

$$\begin{aligned} d\text{cov}^2(X,Y) = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left[||X-X'||_2||Y-Y'||_2\right] \\ + \mathbb{E}_X\mathbb{E}_{X'}||X-X'||_2 \cdot \mathbb{E}_Y\mathbb{E}_{Y'}||Y-Y'||_2 \\ - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}||X-X'||_2 \cdot \mathbb{E}_{Y'}||Y-Y'||_2\right] \end{aligned}$$

X' and Y' are independent copies of X and Y respectively. $d\text{cov}^2(X,Y)$ is 0 if and only if X and Y are independent and non-zero otherwise (Szekely, 2007) (Sejdinovic, 2013). We can also represent this function as

$$d\text{cov}^2(X,Y) = \int_{\mathbb{R}^p \times \mathbb{R}^q} |\mathbb{E}[g_{XY}(u,v) - \mathbb{E}[g_X(u)]\mathbb{E}[g_Y(v)]]|^2 w(u,v)dudv$$

where $\mathbb{E}[g_{XY}(u,v)] = \mathbb{E}[e^{u^TX+v^TY}]$ is the joint characteristic function of (X,Y), while $\mathbb{E}[g_X(u)] = \mathbb{E}[e^{u^TX}]$ and $\mathbb{E}[g_Y(v)] = \mathbb{E}[e^{v^TY}]$ represent the marginals. It is clear in this representation that Independence would force the joint characteristic function to be equal to the product of the marginals, setting the integral to 0.

This function can be normalized to the *distance correlation* function $d\text{corr}$ as

such:

$$d\text{corr}^2(X,Y) = \begin{cases} \frac{d\text{cov}^2(X,Y)}{\sqrt{d\text{cov}^2(X,X)d\text{cov}^2(Y,Y)}} & d\text{cov}^2(X,X)d\text{cov}^2(Y,Y) > 0 \\ 0 & d\text{cov}^2(X,X)d\text{cov}^2(Y,Y) = 0 \end{cases}$$

$d\text{corr}(X,Y)$ is bounded between 0 and 1, similar to Pearson's correlation. Unlike the case of Pearson's correlation, however, dependent but uncorrelated random variables will have non zero values for $d\text{cov}^2(X,Y)$ and $d\text{corr}^2(X,Y)$. Take for example, $X \sim \mathcal{N}(0,1)$ and $Y = X^2$. $\text{Cov}(X,Y) = 0$ while $d\text{corr}(X,Y) \approx 0.782$. Therefore, distance covariance and correlation make a desirable measure of both linear and nonlinear dependence.

Another common measure of dependence is the Hilbert-Schmidt Information Criterion (HSIC) whose estimator operates on the kernel matrices K_X and K_Y of the sample $\{(X_i, Y_i)\}_{i=1}^n$. These methods are equivalent due to a bijective mapping between distance functions and kernels shown in Shen and Vogelstein, 2018. We focus here on distance-based methods. Our task will be to apply this measure to a joint time series process.

4.2 Empirical estimate of $d\text{cov}^2(X,Y)$ and $d\text{corr}^2(X,Y)$

Given sample $\{(X_1,Y_1),...,(X_n,Y_n)\}$, the empirical estimate $d\text{cov}_n^2(X,Y)$ is computed as follows. Generate two $n \times n$ distance matrices $[a_{ij}] = ||X_i - X_j||_2$ and $[b_{ij}] = ||Y_i - Y_i||_2$, respectively. Double center the matrices $[a_{ij}]$ and $[b_{ij}]$ so that their column and row means are both zero. This yields matrices A and

B:

$$A_{ij} = a_{ij} - \overline{a}_{i\cdot} - \overline{a}_{\cdot j} + \overline{a}_{\cdot\cdot}$$

$$B_{ij} = b_{ij} - \overline{b}_{i\cdot} - \overline{b}_{\cdot j} + \overline{b}_{\cdot\cdot}$$

where $\overline{a}_{i\cdot} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, $\overline{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n-j} a_{ij}$, and $\overline{a}_{k\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n-j} a_{ij}$. The notation is analogous for $\overline{b}_{i\cdot}$, $\overline{b}_{\cdot j}$, $\overline{b}_{\cdot\cdot}$. Finally, compute:

$$d\text{cov}_n^2(X,Y) = \frac{1}{n^2}\sum_{i=1,j=1}^{n} A_{ij}B_{ij}$$

This is a biased estimate for $d\text{cov}^2(X,Y)$. An unbiased estimate is given by Shen and Vogelstein, 2018. For the unbiased estimate to be well-defined, we must have $n > 3$. Let $d\text{cov}_{n\text{U}}^2(X,Y)$ denote this estimator.

$$d\text{cov}_{n\text{U}}^2(X,Y) = \frac{1}{n(n-3)}\sum_{i\neq j}\tilde{A}_{ij}\tilde{B}_{ij}$$

For distance matrices $\tilde{A}$ and $\tilde{B}$, with $\tilde{A}_{ij}$ defined below ($\tilde{B}_{ij}$ is analogous).

$$\tilde{A}_{ij} = \begin{cases} a_{ij} - \frac{1}{n-2}\sum_{j=1}^{n} a_{ij} - \frac{1}{n-2}\sum_{i=1}^{n} a_{ij} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^{n} a_{ij}, & i \neq j \\ 0, & i = j \end{cases}$$

Finally, an important note is that while $d\text{corr}^2$ is a well interpretable measure of dependence, its sample equivalent may not be well-suited as an ad-hoc measure in place of a formal test. Indeed, while this may work for univariate X and Y, Dueck et all. (2014) showed that for $(X,Y) \in \mathbb{R}^{p+q}$ with fixed q,

$$\lim_{p\to\infty} d\text{corr}(X,Y) = 0$$

regardless of the dependence structure between X and Y at any p. Szekely et al. (2013) showed that for fixed n,

$$\lim_{p,q\to\infty} d\text{corr}_n(X,Y) = 1$$

if the components of X and Y are i.i.d. and second moments exist. These effects complicate the interpretability of $d\text{corr}_n$, which is why it is best to estimate its sampling distribution. As a result, whether normalized or unnormalized, $d\text{cov}_n^2$ and $d\text{corr}_n^2$ function more appropriately as test statistics rather than estimators (Edelmann and Pitsillou, 2018).

4.3 Multiscale Graph Correlation (`MGC`)

`MGC` builds on the distance correlation test statistic by retaining only the distances that are most informative toward the relationship between X and Y. Specifically, let A and B be the double-centered distance matrices above. Define G_k and H_l to be the k-nearest and l-nearest neighbor matrices, respectively. $[G_k]_{ij} = 1$ indicates that A_{ij} is within the smallest k values of the i-th row of A, and similarly for H_l (Vogelstein and Shen, 2019). Define:

$$d\text{cov}_n^{kl}(X,Y) = \sum_{i,j} A_{ij}[G_k]_{ij}B_{ij}[H_l]_{ij}$$

Each value of $d\text{cov}_n^{kl}$ is normalized by dividing by $\sqrt{\sum_{i,j} A_{ij}^2[G_k]_{ij} \times \sum_{i,j} B_{ij}^2[H_l]_{ij}}$. The final test statistic $\text{mgc}_n(X,Y) = \left\{\max_{k,l} d\text{cov}_n^{kl}(X,Y)\right\}$ is the smoothed maximum of the $\{c_{kl}\}$ over k and l, giving this statistic better finite-sample performance, theoretical guarantees, and mitigating bias (Vogelstein and Shen, 2019).

4.4 Linear dependence in time series

For a stationary time series, $\{X_t\}_{t=-\infty}^{\infty}$, linear dependence is measured with the *autocovariance* function (ACVF) at lag $j \in \mathbb{N}$.

$$\text{ACVF}(j) = \text{Cov}(X_t, X_{t+j})$$

The normalized *autocorrelation* function (ACF) is defined as:

$$\text{ACF}(j) = \frac{\text{ACVF}(j)}{\text{ACVF}(0)}$$

To compute the empirical estimates, generate the samples $\{X_1, ..., X_{n-j}\}$ and $\{X_{j+1}, ..., X_n\}$, and compute the sample covariance and correlation by standard methods.

Analogously, given two stationary time series $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$, the *crosscovariance* function (CCVF) at lead/lag $j \in \mathbb{Z}$ is defined as:

$$\text{CCVF}(j) = \text{Cov}(X_t, Y_{t+j})$$

The normalized *crosscorrelation* function (CCF) is defined as:

$$\text{CCF}(j) = \frac{\text{CCVF}(j)}{\sqrt{\text{ACVF}_{X_t}(0)\text{ACVF}_{Y_t}(0)}}$$

This measure pairwise linear dependence between pairs of components from either time series. The sample equivalents are computed by generating $\{X_1, ..., X_{n-j}\}$ and $\{Y_{j+1}, ..., Y_n\}$ and taking the sample covariance and correlation. Plotting the estimates of this function is usually the first resort for a researcher wishing to investigate the relationships within observations (ACF) or between observations (CCF). A result due to Bartlett gives 95% confidence

bands $\pm\frac{1.96}{\sqrt{n}}$ for these estimates, and statistical programs such as R automatically overlay them on ACF and CCF plots. However, this formula applies for linear time series, as in $X_t = \mu + \sum_{i=-\infty}^{\infty} \psi_i Z_{t-i}$ where $\mu, \phi_j \in \mathbb{R}$ and $Z_j \sim \mathcal{N}(0, \tau^2)$ for $\tau^2 > 0$. For nonlinear models, these bands can be uninformative, and ad-hoc analysis of the ACF or CCF plot is potentially misleading (Politis, 2003). For this reason, we are interested in estimating a function similar to the CCVF/CCF that will capture nonlinear dependencies.

4.5 Other dependence measures in time series

The *standardized spectral density* of a stationary time series is defined as

$$h(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \text{ACF}(j) e^{-ij\omega},$$

i.e. the Fourier transform of the ACF. In white noise settings, i.e. the time series in uncorrelated with itself at various lags, $h(\omega)$ is uniform, as all frequencies are represented equally in the spectrum. Deviation from uniformity implies autocorrelatedness of the time series. Similarly, Hong (Hong, 1999) defines the *generalized spectral density*

$$f(\omega, u, v) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j(u, v) e^{-ij\omega}$$

with $\sigma_j(u, v) = |\mathbb{E}[g_{X_t, X_{t-j}}(u, v)] - \mathbb{E}[g_{X_t}(u)]\mathbb{E}[g_{X_{t-j}}(v)]|^2$. Deviations of the generalized spectral density from uniformity imply independence (Hong, 1999). To generate a test statistic based on this observation, (Fokianos and

Pitsillou, 2017) uses:

$$\int_{-\pi}^{\pi} ||\hat{f}_n(\omega, u, v) - \hat{f}_0(\omega, u, v)||_w^2 d\omega$$

where $\hat{f}_n(\omega, u, v)$ is an estimate of the generalized spectral density, and $\hat{f}_0(\omega, u, v) = \frac{1}{2\pi}\hat{\sigma}_0(u, v)|$ is an estimate of the uniform density under the assumption of independence, and $||\cdot||_w$ is a weighted L_2-norm with respect to some weight function $w(u, v)$. Fokianos and Pitsillou, 2017 gives kernel-type estimators for the spectrum (as is common in spectral estimation (Cryer and Chan, 2011)), and using the weight function described in 4.1, the statistic can be written as:

$$\int_{-\pi}^{\pi} ||\hat{f}_n(\omega, u, v) - \hat{f}_0(\omega, u, v)||_w^2 d\omega = \frac{2}{\pi} \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) k^2 \left(\frac{j}{p}\right) d\text{cov}_n(X_t, X_{t-j})$$

where $k(\cdot)$ is a kernel function, and p is a bandwidth parameter. (Fokianos and Pitsillou, 2017) offer consistency results for a test of serial dependence based on this test statistic, when $p = cn^\lambda$ for $c > 0$ and $\lambda \in (0, 1)$. For this reason, we can adapt the methodology for testing for dependence between stationary time series.

5 Methodology

Recall the problem statement in Section 3, that is given $\{(X_1, Y_1), ..., (X_n, Y_n)\} \sim F_{\{(X_t, Y_t)\}}$, we test:

$$H_0 : F_{(X_t, Y_{t-j})} = F_{X_t} F_{Y_{t-j}} \text{ for all } j \in \{0, 1, ..., M\}, \ \forall t \in \mathbb{Z}$$

$$H_A : F_{(X_t, Y_{t-j})} \neq F_{X_t} F_{Y_{t-j}} \text{ for some } j \in \{0, 1, ..., M\}, \ \forall t \in \mathbb{Z}$$

The test comprises of a parameter as a measure of dependence, its estimator used to derive a test statistic, and resampling procedure as a way to estimate the p-value of the observed test statistic.

5.1 The `DCorr-X` and `MGC-X` independence tests

Define the *cross-distance correlation* parameter at lag j as:

$$d\text{cov}^2(j) := d\text{cov}^2(X_t, Y_{t-j})$$

Where $d\text{cov}^2(\cdot, \cdot)$ is the distance correlation function described in section 4.1. Assuming strict stationarity of $\{(X_t, Y_t)\}$ is important in even defining $d\text{cov}^2(j)$, as the parameter depends only on the spacing j, and not the timestep t of X_t and Y_{t-j}. Similarly, let $d\text{cov}_n^2(j)$ be its estimator. The `DCorr-X` test statistic T_n is defined as:

$$T_n^{(M)} = \sum_{j=0}^{M} \left(\frac{n-j}{n} \right) \cdot d\text{cov}_n(X_t, Y_{t-j})$$

The `MGC-X` test statistic T_n is defined as:

$$T_n^{(M)} = \sum_{j=0}^{M} \left(\frac{n-j}{n} \right) \cdot \text{mgc}_n(X_t, Y_{t-j})$$

`MGC-X`, while more computationally intensive, employs multiscale analysis to achieve better finite-sample power (Vogelstein and Shen, 2019).

Finally, to estimate the distribution of $T_n^{(M)}$ given above, a bootstrap procedure for dependent data is required. A typical block bootstrap captures the dependence between elements of the series. The fixed size bootstrap is employed in the simulations in Section 6, with $b = \sqrt{n}$. Specifically,

1. Uniformly sample (approximately) $\frac{n}{b}$ indices, $\{t_1, t_2, ..., t_{\frac{n}{b}}\}$ selected from $\{1, ..., n\}$.

2. From index t_i, produce block $B_i = (Y_{t_i}, Y_{t_i+1}, ..., Y_{t_i+b-1})$.

3. Let the series $\{Y_\pi(1), ..., Y_\pi(n)\} = (B_1, B_2, ..., B_{\frac{n}{b}})$, where π maps indices $\{1, 2, ..., n\}$ to the new bootstrapped indices.

4. Compute T_n^* on the series $\{(X_t, Y_t^*)_{t=1}^n\}$

Repeat this procedure B times (typically $B = 100, 1000$), and let the α-th critical value of T_n be the top α-th percentile of this empirical distribution. This a yields a critical value and p-value for the test.

5.2 Estimating the optimal lag

The researcher might wish to know value of lag j that maximizes the dependence between X_t and Y_{t-j} (the "optimal lag"), giving more information as to the nature of the relationship between the two time series. Denote this M^*. Both procedures yield estimator $\hat{M}$.

$$\hat{M} = \begin{cases} \arg\max_j \left(\frac{n-j}{n}\right) \cdot d\text{cov}_n(X_t, Y_{t-j}) \\ \arg\max_j \left(\frac{n-j}{n}\right) \cdot \text{mgc}_n(X_t, Y_{t-j}) \end{cases}$$

6 Simulation Results

6.1 Power Curves

The functions to perform the test are implemented in the `pip` installable Python package `mgcpy`. The following simulations are run $R = 100$ times, with

$B = 100$ bootstrap replicates, at $\alpha = 0.05$. The power is estimated at varying sample sizes. ϵ_t represents the noise on time series $\{X_t\}$ and η_t represents the noise on time series $\{Y_t\}$, both generated as standard normal random variables.

6.1.1 Example: Validity against independence

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix}$$

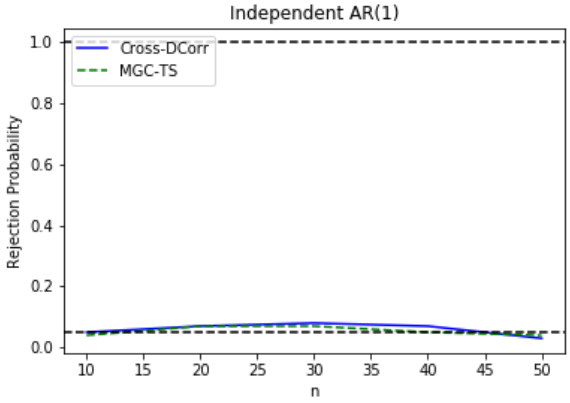


Figure 1: While the block bootstrap procedure is known to have results in terms of **asymptotic** validity, simulations show validity at sample sizes as low as 50.

6.1.2 Example: Consistency against correlation

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix}$$

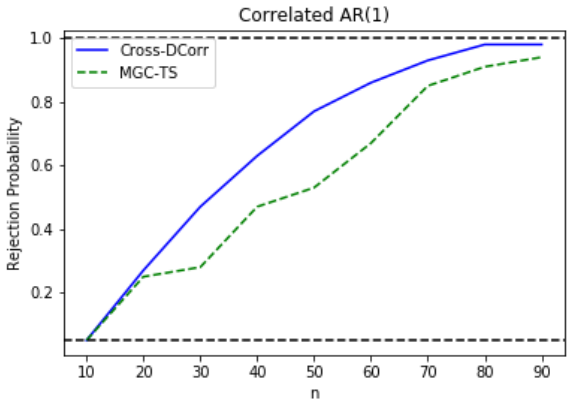


Figure 2: In the linear case, `DCorr-X` (always) chooses the global scale while `MGC-X` might choose a local scale due to random variation.

6.1.3 Example: Consistency against nonlinear dependence

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \epsilon_t Y_{t-1} \\ \eta_t \end{bmatrix}$$

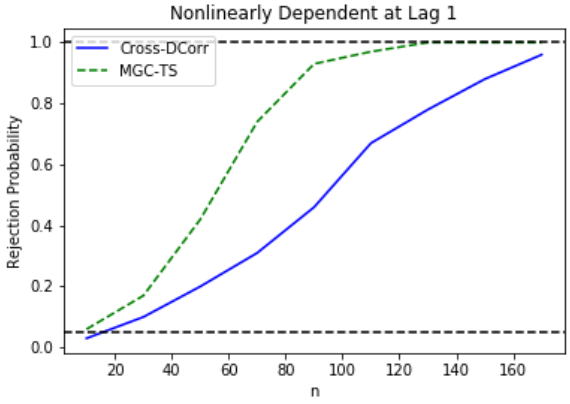


Figure 3: Finally, `MGC-X` enjoys high finite-sample power in nonlinear settings, while `DCorr-X` converges, but significantly slower.

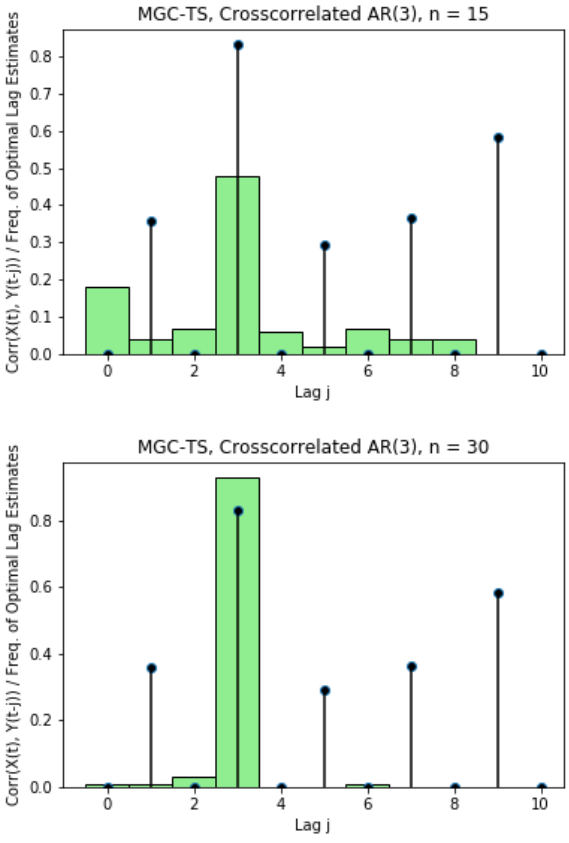
`MGC-X` and `DCorr-X` are both consistent, with the optimal scale of `MGC-X` increasing the finite sample power significantly for the nonlinear setting.

6.2 Optimal Lag Estimation

We delve deeper into estimation of the optimal lag, i.e. the lag time in which the highest dependence exists between X_t and Y_{t-j}. In the following linear time series simulations, the stems correspond to the true cross-correlation, CCF(j), while the bars represent the empirical distribution of the optimal lag estimate for $R = 100$ trials.

Consider the process:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} 0 & \phi_1 \\ \phi_1 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} 0 & \phi_3 \\ \phi_3 & 0 \end{bmatrix} \begin{bmatrix} X_{t-3} \\ Y_{t-3} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix}$$



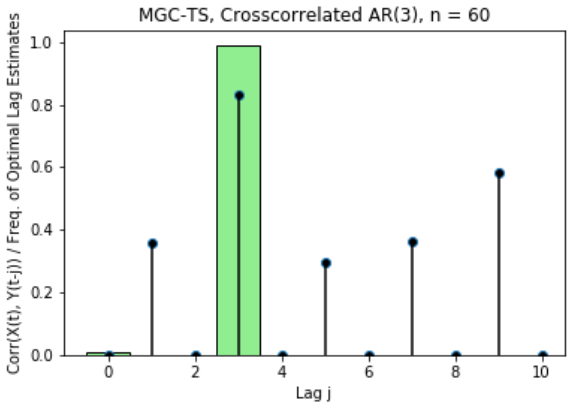
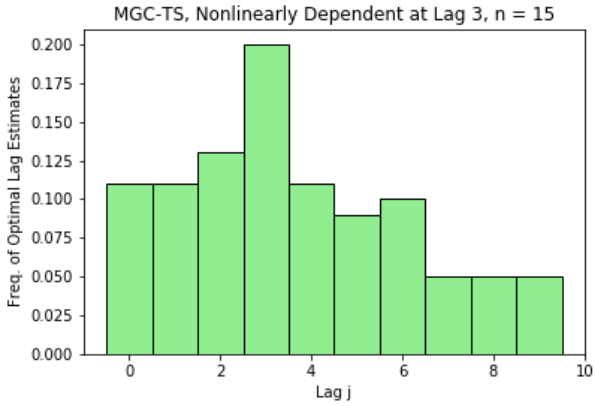


Figure 4: $\phi_1 = 0.1, \phi_3 = 0.8, \epsilon_t, \eta_t \sim \mathcal{N}(0,1)$. For $\phi_3 >> \phi_1$, it is clear that there exists a strong dependence between the X_t and Y_{t-3}. `DCorr-X` and `MGC-X` close in on the correct lag as n grows.

Similarly, consider the nonlinear process, which has clear dependence at lag 3:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \epsilon_t Y_{t-3} \\ \eta_t \end{bmatrix}$$



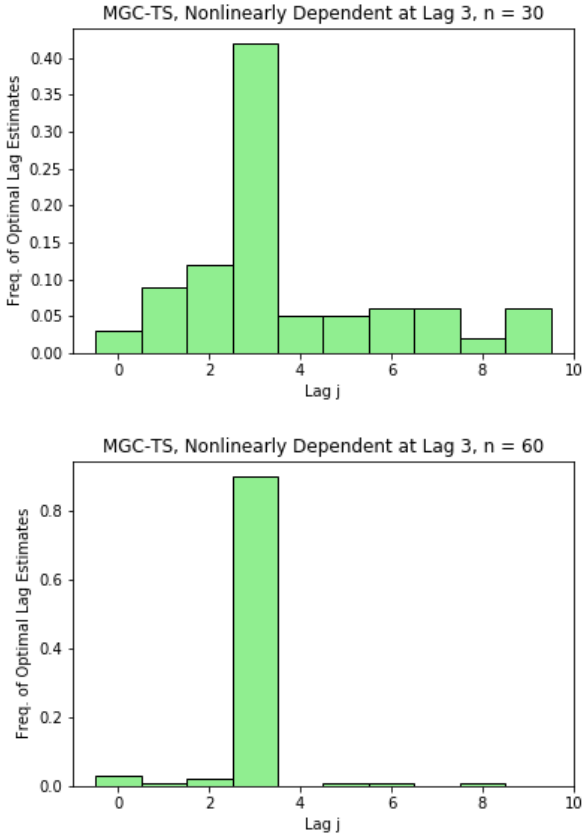


Figure 5: We observe a similar phenomenon in the distribution of optimal lag estimates, as center more tightly around the true optimal lag $j = 3$ for even the relatively small sample size of $n = 60$.

7 Conclusion

The results regarding `DCorr-X` and `MGC-X` prompt further analysis. Work in progress includes theoretical analysis, extending the work of (Shen, 2018) in the dependence setting. On the applications side, these methods motivate research into biological time series, for which finite-sample testing power is especially crucial. Methodological extensions include K-sample testing of time series in light of (Shen and Vogelstein, 2018). In any case, researchers

of many disciplines now have a first resort data analysis tool that deciphers hidden relationships in time series, and further expands the reach of statistical principles in making scientific discovery possible.

Data and Code Availability Statement The analysis of the data was performed using an open-source software package MGCPy (https://mgcpy.neurodata.io).

References

Vogelstein Joshua, Eric Bridgeford Qing Wang Carey E. Priebe Mauro Maggioni and Cencheng Shen (2019). "Discovering and Deciphering Relationships Across Disparate Data Modalities". In: *ELife*.

Behrens, Timothy EJ and Olaf Sporns (2012). "Measuring nonlinear dependence in timeâĂŘseries, a distance correlation approach". In: *Journal of Time Series Analysis* 33, pp. 438–457. DOI: 10.1080/10618600.2016.1193505. URL: https://projecteuclid.org/euclid.ss/1063994977.

Gretton, Arthur (2005). "Measuring Statistical Dependence with Hilbert-Schmidt Norms". In: *International Conference on Algorithmic Learning Theory*. URL: https://link.springer.com/chapter/10.1007/11564089_7.

Shen, Cencheng and Joshua T. Vogelstein (2018). "The Exact Equivalence of Distance and Kernel Methods for Hypothesis Testing". In: *CoRR abs/1806.05514*.

Wang Guochang, Wai Keung Li and Ke Zhu (2018). "New HSIC-based Tests for Independence between Two Stationary Multivariate Time Series". In: *ArXiv E-prints*.

Szekely Gabor J, et al. (2007). "Measuring and Testing Dependence by Correlation of Distances". In: *The Annals of Statistics*. URL: projecteuclid.org/euclid.aos/1201012979.

Sejdinovic, Dino (2013). "Equivalence of Distance-based and RKHS-based Statistics in Hypothesis Testing". In: *Annals of Statistics* 41.5.

Edelmann Dominic, Konstantinos Fokianos and Maria Pitsillou (2018). "An Updated Literature Review of Distance Correlation and Its Applications to Time Series". In: *International Statistical Review*.

Politis, Dimitris N (2003). "The Impact of Bootstrap Methods on Time Series Analysis". In: *Statist. Sci.* 18.2, pp. 219–230.

Hong, Yongmiao (1999). "Hypothesis Testing in Time Series via the Empirical Characteristic Function: A Generalized Spectral Density Approach". In: *Journal of the American Statistical Association* 94.448, p. 1201. DOI: 10.2307/2669935.

Fokianos, K. and M. Pitsillou (2017). "Consistent Testing for Pairwise Dependence in Time Series". In: *Technometrics* 59.2.

Cryer, Jonathan D. and Kung sik Chan (2011). "14.9: Other Methods of Spectral Estimation.âĂİ Time Series Analysis with Applications in R". In:

Shen Cencheng, et al (2018). "From Distance Correlation to Multiscale Graph Correlation". In: *Journal of the American Statistical Association*, pp. 1–22. DOI: 10.1080/01621459.2018.1543125.

Biographical Note

I was born in 1997 in the United States, and have completed my Bachelor's and Master's degree in Applied Mathematics and Statistics at Johns Hopkins University. From my very first course, I grew an immediate passion for mathematical statistics. Given the precision and elegance of mathematics with far-reaching applications to real-world data, this was the subject that would define my education. My goal is to be an informed engineer and mathematician - to understand problems from a theoretical perspective as well as an applied one.

During my 2017 summer at the JHU Applied Physics Laboratory, I tackled network and sentiment analysis of Twitter data, as well as unsupervised learning for categorical data. At Goldman Sachs my next summer, I learned a proprietary language during my internship, and used it to implement software for analysis of time series data.

Throughout my time during the academic years, I have worked extensively with Dr. Joshua T. Vogelstein in the Department of Biomedical Engineering. My projects included a visualization application for high-dimensional, noisy brain images, an `R` package for statistical inference on graphs, and this work. My current career objectives are to pursue a Ph.D., enter the exciting and fast-growing field of machine learning research, and continuing learning for the rest of my life.

Ronak Mehta